0

SZIMBÓLUMSOROZATOK ELEMZÉSE STATISZTIKAI MÓDSZEREKKEL

Tudományos diákköri dolgozat, 1997.

Készítette: **Bíró Tamás**
V. éves fizikushallgató
ELTE TTK

Témavezető: **Vicsek Tamás**
egyetemi tanár
ELTE TTK Atomfizikai Tanszék

Budapest, 1997.

Összefoglaló

Dolgozatom célja az, hogy összefoglalja azokat a statisztikus fizikai és sztochasztikus módszereket, amelyek szimbólumsorozatok (szövegek természetes nyelveken, DNS-kód,...) elemzésénél hatékonyan használhatóak. Beketintést adok az ezen a téren a kilencvenes években végzett kutatásokba (pl. hosszútávú korrelációk, Zipf-törvény). Majd megvizsgálom, milyen következményekkel járnak ezen eredmények a szimbólumsorozatot létrehozó folyamatok modellezhetőségére nézve, ha a formális nyelvek elméletét kibővítjük sztochasztikus eszközökkel. Azt kapom, hogy a folyamatot két szakaszra kell bontanunk, az első, amely egy nem-lineáris "kódolandóból" készít egy lineáris kódot, hosszútávú korrelációkat eredményez; viszont a második szakasz elemezhető Markov-folyamatként. Bevezetek egy eljárást, amely szekvenciákat hasonlít össze, a rövidtávú korrelációkból adódó jellegzetességeik alapján, és felhasználom ezt a módszert a fonotaktikában, valamint kódoló és nem-kódoló DNS-szakaszok összehasonlítására.

Tartalomjegyzék

1. Bevezetés

Míg a hagyományos fizika az egyes tömegpontok, részecskék viselkedésével foglalkozik, a modern statisztikus fizika a nagyszámú kölcsönható részecske által létrehozott *struktúrákat* is képes leírni. Ezek a struktúrák gyakran függetlenek az őket létrehozó részecskék legtöbb speciális tulajdonságától, ezt nevezzük *univerzalitásnak*.

Kölcsönható "részecskékből" álló rendszerekkel a fizikán kívül is találkozunk. Az elmúlt években sikerrel alkalmazták a statisztikus fizikában kifejlesztett módszereket a molekuláris biológiában [1], az evolucióbiológiában [2], vagy éppen a közdazdaságtanban, [3] ahol a piacon versengő vállalatok vagy a tőzsdén "kölcsönható" ügynökök a statisztikus fizikából ismert viszonyokat hoznak létre. Az egyensúlytól távoli rendszerekre jellemző skálázási (ill. fraktális) tulajdonságokkal (Vicsek [1992], Family & Vicsek [1991]) találkozunk például a vállalatok méret szerinti eloszlásánál ,[4] de előbukkannak statisztikus fizikai módszerek az opciók árazásánál és a tőzsdei árfolyamok alakulásánál is [5].

Az emberi beszéd vizsgálatánál szintén találkozunk "kölcsönható rendszerekkel", méghozzá két szinten is. Az állati kommunikációtól ugyanis éppen az ún. "kettős tagoltság" ("kettős szerkezet") különbözteti meg az emberi beszédet (Kenesei [1995], p. 31.): a jelentés nélküli elemi hangok ("fonémák", a klasszikus nyelvészeti iskolák szerint) építik fel a szavakat ("morfémákat"), melyekből pedig a mondatok ("megnyilatkozások") épülnek fel. Mind a két szinten kölcsönhatnak a magasabb egységet alkotó építő elemek. A hangok rendszerét és kölcsönhatásait írja le a fonológia [6], míg a morfémák kölcsönhatásait a szintaxis (azaz "mondattan"). Joggal merülhet fel a kérdés, hogy vajon a nyelvi rendszer(ek) is rendelkezik (rendelkeznek)-e az említett univerzális tulajdonságokkal. Az elmúlt évek vizsgálatai azt mutatják, mint azt a II. fejezetben ismertetni fogom, hogy igen.

---

[1] Ld. pl. Derényi & Vicsek [1996].

[2] Ld. pl. Geritz et al. [1997]

[3] Több cikk is található a komplex rendszerek és a közgazdaságtan kapcsolatáról Martinás & Moreau [1995]-ben: pl. J. A. Holyst et al.: Control of Microeconomical Chaos.

[4] M. H. R. Stanley et al. [1996a,b], Amaral et al. [1997]

[5] Bouchaud & Potters [1997], Potters et al. [1997], Ghashghaie et al. [1996], Mantegna & Stanley [1996], Liu et al. [1997], Mantegna & Stanley [1995b, 1996]

[6] A fonológia nem keverendő össze a fonetikával, a hagyományos értelemben vett hangtannal, mely a beszédhangok fizikai tulajdonságaival, valamint képzésük és érzékelésük fiziológiájával foglalkozik.

A természetes nyelvi szövegeken, pontosabban fogalmazva *írott* szövegeken végzett vizsgálatokkal analóg vizsgálatokat lehet végezni DNS-szekvenciákon. A vizsgálati módszer szempontjából a különbség csupán abban az alapábécében van, melyből felépül a szimbólumsorozat. Több, fizikusok által írt cikk jelent meg az elmúlt években, melyek e két szekvenciatípus hasonlóságaira mutatnak rá [7]. Viszont a DNS "nyelve" vagy "nyelvei" (amennyiben beszélhetünk egyáltalán ilyenről, ld. Mantegna et al. [1994, 1995a] megjegyzései) sokkal kevésbé ismert, jóval nehezebben megismerhető, mint a természetes nyelvek rendszerei. Ugyanis sokat segít az utóbbi megismerésében a természetes nyelvek grammatikájáról mindannyiunkban lévő intuíció.

Tehát amennyiben sikerül a nyelvi intuíciónk segítségével felállított nyelvészeti elméleteket a statisztikus fizika eszközeivel is megvizsgálni, a DNS statisztikai tulajdonságai elvezethetnek a genetikai kód grammatikájának jobb megismeréséhez. Talán néhány, jelenleg még megmagyarázatlan tény és jelenség értelmezéséhez is közelebb kerülhetünk. Az egyik legnagyobb ilyen kérdés a DNS nem kódoló szakaszainak, például az *intronoknak*, a szerepére vonatkozik. Lehet, hogy ezek a szakaszok is fontos információkat tartalmaznak, mások szerint "biztonsági" jelentőségük van (növelik a genetikai kód redundanciáját), de létezik olyan vélemény is, mely szerint semmi jelentőségük sincs, csupán "evolúciós szemetek", a múlt emlékei.

Az alkalmazott eljárások lényege az, hogy – a fizika jellegének megfelelően – a jelenségeket kvantitatív "síkra tereljük". De a felhasznált módszerek jóval összetettebbek, mint a hagyományos *kvantitatív nyelvészet* [8] statisztikai módszerei. A kapott eredményeket, a természettudományos megismerés szabályai szerint, modellekkel igyekszünk összevetni. A jelen esetben sztochasztikus eszközök jöhetnek szóba, úgy mint a Markov-láncok és a formális nyelvek elméletének sztochasztikus formalizmussal történő kiterjesztései. Ehhez kapcsolható majd a későbbiekben a szintén fizikai fogalmak analógiájára született, Shanonféle információelmélet.

A 2. fejezetben összefoglalom az 1990-es évek első felében DNS-szekvenciákon és írott nyelvi szövegeken végzett, a szimbólumszekvenciák hosszútávú korrelációinak feltárására irányuló kutatások eredményeit. A továbbiakban arra leszek kíváncsi, hogy ezek az eredmények milyen következményekkel járnak a kódoló mechanizmusra nézve. Feltételezem, hogy a szimbólumsorozat leírható a formális nyelvek eszközeivel, amely elméletnek az alapjait a 3. fejezetben ismertetem. A 4. fejezetben analógiákra mutatok rá a DNS- és a természetes nyelvi jelsorozatok között. Az 5. fejezetben kibővítem a formális nyelvek elméletét sztochasztikus eszközökkel, hogy az általam javasolt analógiák segítségével megmagyarázhassuk a két típusú szekvencia hasonló tulajdonságait. A 6. fejezetben bevezetek egy olyan módszert, amely, az írott szövegek rövidtávú korrelációkból adódó jellegzetességei alapján, a szövegek nyelvi és tartalmi hasonlóságára ad egy "mértéket". Ezt a módszert a 7. fejezetben DNS-szekvenciákra is alkalmazom, és rámutatok a kódoló és nem-kódoló szakaszok újabb eltérő viselkedésére. Végezetül, a 8. fejezetben összefoglalom eredményeimet.

[7] Dietler & Zhang [1994], Mantegna et al. [1994, 1995a].

[8] Példaként ld. Nagy F. [1986]-ot.

2. Hosszútávú korrelációk

A kilencvenes években egyre nagyobb mennyiségben készülnek el virális, bakteriális, növényi, állati és emberi kromoszómák fizikai térképei. Ezen adatok meglehetősen "unalmas" adatbázisokban találhatók meg: a négy bázisnak (`A`: adenin, `C`: citozin, `G`: guanin és `T`: timin) megfelelően, négy betű véget nem érő sorozatát alkotják. Időnként változatosságot jelent egy-egy egzotikus bázis felbukkanása. Miféle érdekes információ nyerhető ezen százezres nagyságrendű bázispár-szekvenciából?

1992-ben C.-K. Peng és társai, többségük a Bostoni Egyetem munkatársa, érdekes megfigyeléseiket hozták nyilvánosságra a Nature hasábjain. [1] Egydimenziós bolyongássá átalakítva a DNS-szekvenciát ("DNA walk"), hosszútávú korrelációkat fedeztek fel az intronokat tartalmazó szekvenciákban, melyek hiányoztak az exonokban, valamint a c-DNS-ekben, [2] és az intront nem tartalmazó génekben. A módszert más fizikusok felhasználták írott szövegek elemzésére is. [3]

Az eljárás lényege a következő: képzeljünk el egy bolhát, melyet egy számegyenes origójába helyezünk. Felolvassuk neki a szekvenciát, és amennyiben pirimidinvázas bázist hall (`C` vagy `T`), úgy előre lép egyet, míg purinvázas bázis esetén (`A` vagy `G`) hátra lép. Amennyiben u_i-vel jelöljük az i-ik lépést ($u_i = \pm 1$), úgy a bolha helyzete az i-ik lépés után nyilván

$$y(l) = \sum_{i=1}^{l} u_i. \tag{2.1}$$

A mozgást jellemző statisztikus fizikai mennyiség az $y(l)$-nek az elmozdulás átlaga körül vett $F(l)$ átlagos fluktuációja (*root mean square fluctuation*):

$$F^2(l) = \left\langle (\Delta y(l) - < \Delta y(l) >)^2 \right\rangle = \left\langle \Delta y(l)^2 \right\rangle - \left\langle \Delta y(l) \right\rangle^2, \tag{2.2}$$

ahol

$$\Delta y(l) = y(l_0 + l) - y(l_0), \tag{2.3}$$

és az összes lehetséges l_0 pozícióra kell az átlagolást képezni.

Az $F^2(l)$ szoros összefüggésben áll az

$$C(l) =< u(l_0)u(l_0 + l) > - < u(l_0) >^2 \tag{2.4}$$

autokorrelációs függvénnyel:

---

[1] Peng et al. [1992], Stanley et al. [1992, 1993a,b], Peng et al. [1993] .

[2] A c-DNS a fehérjeszintézis során használt RNS egy másolata, melyet megfordított transkripcióval kaphatunk meg.

[3] Schenkel et al. [1993], Amit et al. [1994], Dietler & Zhang [1994], Ebeling & Neiman [1995].

$$F^2(l) = \sum_{i=1}^{l} \sum_{j=1}^{l} C(j-i). \tag{2.5}$$

Ebből következik, hogy – $C(l)$ viselkedésétől függően – $F(l)$ háromféle viselkedést mutathat:

• Amennyiben véletlen sorozattal állunk szemben, azaz $C(l) = 0$ átlagban (kivéve az $l = 0$ esetet, hiszen $C(0) = 1$), akkkor klasszikus, korrelálatlan bolyongásról van szó: $\alpha = 0,5$ kitevőjű hatványfüggvényként cseng le az $F(l)$.

• Rövidtávú korrelációk esetén adott a korrelációk R karakterisztikus hossza, azaz az autokorrelációs függvény $C(l) \sim \exp(-l/R)$ viselkedést mutat. Az $F(l)$ aszimptótikus viselkedése ebben az esetben is $\alpha = 0,5$ kitevőjű hatványfüggvény lesz.

• Hosszútávú korrelációk esetén viszont, $C(l)$ nem jellemezhető R karakterisztikus hosszal, hanem – exponenciális helyett – hatványfüggvényszerű viselkedést tapasztalunk. Ennek következménye $F(l)$-re nézve az lesz, hogy az α-kitevő értéke eltér a $0,5$ értéktől. Az eltérés mértéke jellemzi a hosszútávú korrelációk erősségét. Az $\alpha = 1$ eset felel meg a skálainvariáns $1/f$ zajnak ("*maximal complexity limit*") (Amit et al. [1993]). Az alábbi esetekben $0,5$ és 1 közötti értékeket fogunk találni.

Az eljárás átvihető természetes nyelvi írott szövegekre is. Ebben az esetben egy-az-egyhez kódolást alkalmaztak, azaz átírták a szöveget egy bináris ábécé segítségével (szemben a "DNA-walk"-kal, ahol is két-két bázist ugyanúgy kódoltak). Például öt biten kitűnően lehet kódolni az angol ábécé 26 betűjét, a szóközt, és a legfontosabb írásjeleket.

Az eredmények nagyon érdekesek. A vizsgált DNS-szekvenciák nagyon jól szétválaszthatók két csoportba: a tisztán kódoló szakaszokból álló szekvenciák $\alpha = 0,50 \pm 0,01$ exponenssel jellemezhetők, azaz nem mutatnak hosszútávú korrelációkat, szemben az intront is tartalmazó szekvenciákkal, melyek esetében $\alpha = 0,61 \pm 0,03$. (Peng et al. [1992])

Írott szövegek esetében, a vizsgálathoz használt szövegek között megtaláljuk a Biblia eredeti, héber nyelvű szövegét, valamint modern fordításait, továbbá Shakespeare-drámákat, regényeket, sőt egy szótárat is. Néhány érdekesebb eredmény:

• A szövegek sok nagyságrenden keresztül állandó α-exponenssel jellemezhetőek, melynek értéke általában $0,6 - 0,7$ közé esik (Schenkel et al. [1993]).

• A kitevő értéke nem jellemző a szerzőre, hiszen például a *Hamlet* exponense $0,56$, míg a *Rómeó és Júliáé* $0,6$ (Schenkel et al. [1993]).

• A fordítások során, úgy tűnik, a korreláció mértéke csökken. Habár a Biblia exponense kimagaslóan magas ($\sim 0,75$), a fordításai során ezen érték szisztematikusan csökken. [4] (Amit et al. [1994])

• A szótár a szócikkek hosszánál jóval hosszabb korrelációkat mutat, amely jelenség nem magyarázható pusztán a tartalom korreláltságával (Schenkel et al. [1993]).

• A szövegeket kisebb részletekre, például szavakra, mondatokra vagy $l = 10 - 10000$ karakter hosszúságú, azonos darabokra szabdalva, a szabdalás hosszának nagyságrendjéig

[4] Nem találtam a szakirodalomban világos választ arra a kérdésre, hogy a magas hatványkitevő miért nem lehet a héber nyelvnek vagy ortográfiának jellegzetes tulajdonsága.

megmaradnak a korrelációk, azon túl pedig eltünnek. Ezek szerint a szótár szócikkeinek elrendezése "nem véletlenszerű", előzetes sejtésünkkel ellentétben, írja (Schenkel et al. [1993]).

Lefordított számítógépprogramokat (`.exe`-file-okat) is megvizsgáltak, ezek esetében $0,9$-et is meghaladó kitevőt találtak (Schenkel et al. [1993]). Érdekes még a "humán véletlenszámgenerátor" vizsgálata: az egyik szerző 0 és 9 között írt le számokat, "véletlenszerűen". A véletlenszerű eloszlásra való törekvés eredményeképpen, rövid távon ($l \sim 10$) antikorrelációt fedezhetünk fel, de ennél hosszabb távon − a számokat generáló személy minden igyekezete ellenére − megjelentek a korrelációk. Ráadásul, a hatványkitevő értéke nem is állandó, tart az 1-hez! A szerzők ezt úgy magyarázzák, hogy − ellentétben a szövegekkel és a számítógépes programokkal −, a véletlenszám generálás nem rendelkezik értelmes céllal, vagyis a tudattalan tényezők, melyeket felelőssé tesznek a hosszútávú korrelációkért, nagyobb szerepet játszhatnak (Schenkel et al. [1993]).

Az alábbiakban azt vizsgálom meg, milyen következményekkel járnak ezek a megfigyelések a jelsorozatot létrehozó folyamatok alkalmas modelljének természetére vonatkozóan. Mindenek előtt, be kell vezetnünk a *formális nyelv* fogalmát (Révész [1979] alapján), melyet széles körben használnak a természetes és számítógépes nyelvek leírására. Amellett is felhozok majd érveket, miért tartom hasznosnak ezen matematikai konstrukció alkalmazásának kipróbálását a genetikában. A formális nyelvek elmélete az algebra bevett ágai közé tartozik, de hasznosnak tartom az alapfogalmak összefoglalását, hiszen ez a témakör nem része a fizikus szak tananyagának.

3. Alapfogalmak a formális nyelvek elméletéből

A *formális nyelvek elméletének* a célja a természetes és számítógépes nyelvek leírása. Az alapgondolat az, hogy a nyelvet, mint a "*jólformált mondatok*"− azaz a nyelvhez tartozó, ábécébeli sztringek − halmazát, nem csak a halmaz elemeinek a felsorolásával, vagy a halmazok megadásánál megszokott, egyszerű szabályok segítségével definiálhatjuk. Megadhatunk egy nyelvet a "szerkezetének" leírásával is, *generatív grammatika* segítségével, valamint olyan automata révén, mely a nyelv "mondatait", "formuláit", és csak azokat fogadja el. Mindkét fogalom egy *véges vagy végtelen* halmazt ad meg, *véges* eszközökkel. Ezen gondolat mögött az áll, hogy a gyermek az anyanyelvét nyilván nem a hallott mondatok egyszerű reprodukálásával sajátítja el: egyrészt, nem hallhatja a nyelv összes, potenciálisan végtelen számú mondatát, és képes olyan mondatot is reprodukálni, amelyet előzőleg nem hallott. ; továbbá, egyszerű ismétlés esetén, nem hibázna, hiszen feltehetőleg csak helyes mondatot hall. Tehát − legalábbis Chomsky és követői szerint − a gyermek a nyelv szabályait sajátítja el.

A formális nyelvek elméletében, valamilyen L *nyelv* egy nemüres, véges Σ halmaz, az ún. *ábécé* fölött, definíció szerint nem más, mint a Σ elemeiből képzett, véges hosszúságú sztringek egy halmaza. (Az n hosszúságú sztring egy rendezett n-est jelent matematikailag.) Jelölje Σ^+ a Σ elemeiből képezett, pozitív (véges) hosszúságú sztringek halmazát, Σ^* pedig

ennek kibővítését az e ("empty") üres, azaz nulla hosszúságú sztringgel. Ekkor egy L nyelv Σ fölött nem más, mint Σ^* egy részhalmaza:

$$L \subseteq \Sigma^*$$

.

Az alábbiakban "szó", "mondat", "jelsorozat", "formula" kifejezéseket a "sztring" szinonímájaként fogom használni, a "jel", "betű", "szimbólum" szavak pedig az ábécé elemeire fognak utalni. A szóhasználat a konkrét implementációtól függ: szintaxis (mondattan) esetén például Σ szavakat tartalmaz.

Két formula, X és Y *konkatenációján* azt a Z formulát értjük, amelyet úgy kapunk, hogy X-et és Y-t "egymás mellé írjuk", összefűzzük: $Z = XY$. Nyilvánvaló, hogy Σ^* a konkatenáció műveletével együtt egységelemes félcsoportot alkot, ahol az e üres szó játssza az egységelem szerepét.

A *generatív grammatika* fogalmának definíciója Révész [1979.] alapján az alábbi:

1. Definíció: Egy G generatív grammatikán az alábbi rendezett négyest értjük: $G = (V_N, V_T, S, F)$, ahol V_N és V_T diszjunkt véges ábécék, $S \in V_N$, az F pedig olyan rendezett (P, Q) pároknak egy véges halmaza, melyekre $P, Q \in V^*$ ($V := V_T \cup V_N$), és P legalább egy V_N-beli jelet tartalmaz.

A V_N elemeit *nemterminálisok* elemeknek, V_T elemeit pedig *terminális* jeleknek fogjuk nevezni, az alábbiakban ismertetendő okokból. Az F elemeit *helyettesítési szabályoknak* (*rewriting rules*) hívjuk, és $P \to Q$ alakban írjuk. Az S kitüntetett nemterminális elem neve: *mondatszimbólum*. Az alábbi módon definiáljuk a *levezetés* fogalmát:

2. Definíció: Adott $G = (V_N, V_T, S, F)$ grammatika esetén, $X, Y \in V^*$-ra azt mondjuk, hogy Y levezethető X-ből, azaz $X \Rightarrow Y$, ha létezik olyan $P_1, P_2, P, Q \in V^*$, hogy $X = P_1 P P_2$, $Y = P_1 Q P_2$, valamint $(P \to Q) \in F$.

Ez azt jelenti, hogy ha X-ben P-t újraírjuk Q-val az egyik F-beli *helyettesítési szabály* alkalmazásával, úgy Y-t kapjuk. Jelöljük a $\Rightarrow$ reláció tranzitív lezártját $\Rightarrow^*$-val, azaz $X \Rightarrow^* Y$ csakkor áll fenn, ha X-ből kiindulva, F-beli újraírøszabályok nulla vagy véges számú alkalmazásával, eljuthatunk Y-ba. Egy G grammatika *által generált $L(G)$ nyelv* az S mondatszimbólumból levezethető, terminális elemekből álló mondatok halmazát jelenti:

$$L(G) := \{P \in V_T^* | S \Rightarrow^* P\}$$

Azt mondjuk, hogy az L nyelvet a G grammatika *generálja*, ha $L = L(G)$. A terminálisok V_T halmazának szerepét a Σ ábécé tölti be, amikor egy nyelvhez grammatikát gyártunk. Két grammatikát *generálás szempontjából ekvivalensnek* nevezünk, ha ugyan azt a nyelvet generálják.

A formális nyelvek elméletének alapvető kérdése az, hogy milyen típusú nyelvekhez milyen grammatikákat adhatunk meg. Azaz milyen következményekkel jár, ha megkötéseket teszünk a helyettesítési szabályok alakjára. *Noam Chomsky* az alábbi *nyelvosztályokat* vezette be (Révész [1979.]):

3. Definíció: A $G = (V_N, V_T, S, F)$ grammatikát i-típusúnak nevezzük, ha az alábbiak közül az i-ik teljesül:

$i = 0$: Nincs semmilyen kikötés.

$i = 1$: Az F minden eleme $Q_1XQ_2 \to Q_1PQ_2$ alakú, ahol $Q_1, Q_2, P \in V^*$, $X \in V_N$ és $P \neq e$, kivéve esetleg az $S \to e$ szabályt, amely viszont csak úgy szerepelhet az F-ben, ha az S nem fordul elő semelyik szabálynak sem a jobb oldalán.

$i = 2$: Az F minden eleme $X \to P$ alakú, ahol $X \in V_N$, és $P \in V^*$.

$i = 3$: Az F minden eleme $X \to PY$ vagy $X \to P$ alakú, ahol $X, Y \in V_N$, és $P \in V_T^*$.

Valamely L nyelvet i-típusúnak mondunk, ha létezik hozzá i-típusú grammatika. Az i-típusú grammatikák osztályát $\mathcal{G}_i$-vel, míg az i-típusú nyelvek családját $\mathcal{L}_i$-vel szokás jelölni. Belátható, hogy ezen nyelvosztályok egymás valódi részhalmazai:

$$\mathcal{L}_3 \subset \mathcal{L}_2 \subset \mathcal{L}_1 \subset \mathcal{L}_0$$

Az $i = 0$ nyelvosztályt *mondatszerkezetű* nyelveknek nevezzük. Az $i = 1$ típusú grammatikák helyettesítési szabályai úgy néznek ki, hogy az X nemterminálist írjuk újra, a $Q_1\_Q_2$ környezetben, ahol $\_$ jelzi X helyét. Ezért az $i = 1$ grammatika-, ill. nyelvosztályt *környezetfüggőnek* (*context-sensitive*, CS) nevezzük, szemben az $i = 2$ *környezetfüggetlen* (*context-free*, CF) osztállyal. Az $i = 3$ esetben pedig *reguláris* osztályokról beszélünk.

A formális nyelvekkel szoros kapcsolatban állnak az *automata-elméletből* ismert gépek. Egy nyelvet elfogad valamely automata, ha a nyelv mondatait, és csak azokat fogadja el. Az egyes nyelvosztályok ilyen alapon megfeleltethetőek az egyes automata-típusoknak. Erre a kérdésre − hely hiányában − nem fogok részletesen kitérni. Egyedül a véges állapotú automaták alapgondolatát ismertetem, mivel a későbbiekben még visszatérek ezek kapcsolatára a Markov-láncokkal.

A reguláris grammatikák csak lokális jelenségeket tudnak leírni. Az alábbiakban látni fogjuk kapcsolatukat a Markov-láncokkal, vagyis az ilyen nyelvek esetén nem várunk hosszútávú korrelációkat (leszámítva természetesen a determinisztikus esetet). Belátható, hogy reguláris nyelvekhez, és csak azokhoz szerkeszthető elfogadó *véges állapotú automata* (Révész [1979.]):

4. Definíció: Egy véges automatán az $A = (K, T, M, q_0, H)$ *rendezett ötöst értjük, ahol*

K *egy véges, nem üres halmaz, az állapothalmaz,*

T *egy véges ábécé, a bemenő ábécé,*

M *a* $K \times T$ *halmaznak egy leképezése a* K*-ra, az átmenetfüggvény,*

$q_0 \in K$ *a a kezdőállapot,*

$H \subseteq K$ *a végállapotok halmaza.*

A véges állapotú automata működése a következő: elindul a kezdőállapotból, az első lépésben beolvassa a *bemenő szalagra* írt mondat első jelét ($\in T$), és a beolvasott jel, valamint az éppen aktuális állapot függvényében, az M által meghatározott állapotba megy át. A következő lépésben újabb jelet olvas be a szalagról, és ez alapján újabb állapotba "ugrik" − feltéve, hogy az aktuális állapot és a beolvasott jel által alkotott rendezett pár eleme az M átmenetfüggvény értelmezési tartományának. És így tovább, egészen addig, amíg el nem akad az automata működése, vagy el nem olvasta az egész mondatot. Azt mondjuk, hogy *az automata elfogadott egy mondatot*, ha végigolvasta azt, és az utolsó

állapot eleme H-nak. (Ez a folyamat formalizálható "állapotsorozatok" definiálásával, de ettől most tekintsünk el.)

A véges állapotú automata jellemzője, hogy nem rendelkezik "memóriával" (végtelen *veremmel*), és ebből következik majd a hosszútávú korrelációk hiánya a reguláris nyelveknél. Például az $L_a := \{a^n b^n | n > 0\}$ nyelv (ahol a^n n darab 'a' konkatenációját jelenti) azért nem lehet reguláris, mert a 'b'-k beolvasásakor "tudnunk kell" azt, hogy mennyi 'a'-t olvastunk be előzőleg. Márpedig, az 'a'-k száma tetszőlegesen nagy lehet, így ezt a végtelen sok lehetőséget nem tudjuk véges sok állapot segítségével "megjegyezni".

Ezzel szemben, a nem-reguláris környezetfüggetlen nyelvek tetszőlegesen sok "beágyazást" tesznek lehetővé. Tipikus példát szolgáltatnak erre – a fentebb megadott L_a halmazon kívül – a számítógépes nyelvek, ahol a ciklusszervezés jelenti az egymásba ágyazott struktúrákat, valamint a zárójelezett matematikai kifejezések. Chomsky [1957] amellett érvel, hogy az angol nyelv – és általában a természetes nyelvek – szintaxisa is környezetfüggetlen grammatikával adható meg, melyet a transzformációk átalakíthatnak. (Ellenpéldaként szokás felhozni egyes holland szerkezeteket, melyeket környezetfüggő szabályokkal érdemes csak leírni. Ezektől azonban tekintsünk el, mint ahogy azt a legtöbb nyelvész is teszi.)

A környezetfüggetlen nyelvek mondatainak jellemzője a látványos hierarchikus szerkezet. Egy formula levezetését egy gráffal (irányított fával) ábrázolhatjuk, melynek "gyökere" a mondatszimbólum, végpontjai ("levelei") a levezetett mondat szavai. A köztes csomópontok pedig a levezetés során megjelenő nemterminálisoknak felelnek meg, amelyből induló élek végpontjait "összeolvasva", azon helyettesítési szabály jobb oldalát kapjuk meg, amellyel újraírtuk a nemterminálist. Így például azt mondhatjuk (nagyon leegyszerűsítve a kérdés nyelvészeti oldalát), hogy az S mondatszimbólumból levezethető magyar mondat áll egy főnévi csoportból (NP), amely az alany szerepét tölti be, és egy igei csoportból (VP). Azaz felírható a következő szabály: S $\rightarrow$ NP VP. Maga az igei csoport is felépülhet igéből (V), tárgyból és határozókból (utóbbiak szintén NP-k, azaz főnévi csoportok): VP $\rightarrow$ VP NP, illetve: VP $\rightarrow$ V. A főnévi csoport pedig állhat főnevekből (N) és melléknevekből (A): NP $\rightarrow$ A NP, és NP $\rightarrow$ N. Majd a V, N és A nemterminálisokat helyettesíthetjük a megfelelő kategóriájú ("szófajú") magyar szavakkal (terminálisokkal). A fenti szabályok példát mutatnak a rekurziót lehetővé tevő szabályokra is, melyek segítségével lehet egy végtelen sok mondatból álló nyelvet leírni a generatív grammatikák véges eszközével.

A környezetfüggetlen nyelvekhez készített elfogadó automaták osztálya az ún. *veremautomaták*. Adott környezetfüggetlen grammatikához készíthető olyan automata is (ún. *parser*), amely a terminálisok lineáris szekvenciájából előállítja az azt létrehozó levezetés(ek) során alkalmazott szabályok sorozatát.

Környezetfüggő nyelvre példa az $L_b := \{a^n b^n c^n | n > 0\}$ nyelv. (A környezetfüggetlen nyelvekre szükséges feltételt kirovó *Bar Hillel lemma* segítségével látható be, hogy L_b-hez nem adható meg környezetfüggetlen grammatika.) A környezetfüggetlenséget kizárja, ha a "korrelációk" egymást keresztezik, nincsennek egymásba ágyazva, mintha megengednénk a következő "zárójelezést":

(...[...)...]

.

Az $\mathcal{L}_1$ nyelvosztály a *lineárisan korlátolt automaták* által elfogadott nyelvek osztályával egyezik meg, míg a mondatszerkezetű nyelveket elfogadó automaták éppen a híres *Turing-gépek*.

Környezetfüggő grammatikát használnak a fonológusok, a nyelvekben lejátszódó hangtani folyamatok jellemzésére. Viszont Kaplan és Kay [1994] bebizonyítja, hogy a fonológiai modellek általában olyan feltételeket rónak ki a szabályok alkalmazására, amelyek regulárissá redukálják a folyamatokat. Erre hivatkozva, az alábbiakban feltételezem, hogy a fonológia − elvileg − felírható reguláris szabályok segítségével is.

Láttuk, hogy a formális nyelvek elmélete éppannyira hasznos a nyelvészetben, mint a számítástechnikában. Úgy vélem, a genetikában is eredményesen lehetne felhasználni ezt a modellt. Hiszen a fehérjék hasonló hierarchikus szerkezettel rendelkeznek, mint a környezetfüggetlen nyelvek. Az elsődleges, másodlagos és harmadlagos szerkezet egymásra épülése, a funkciós csoportok elhelyezkedése, valószínűleg leírható ilyen eszközökkel, de saját magam − hiányos biokémiai ismereteim miatt − nem merek ezen a téren nyilatkozni. De úgy "érzem", hogy a fehérjeszerkezet-kutatások alapján felírhatóak olyan környezetfüggetlen újraíró szabályok, melyeket molekulafizikai számításokkal lehetne igazolni. A gondolatmenetet folytatva, a következő kihívást azt jelentheti, hogy megértsük, ezek a környezetfüggetlen szabályok miként redukálódnak regulárissá, hiszen a c-DNS-ekből hiányzó korrelációk arra engednek következtetni, hogy a fehérjék "szintaxisa" nem csak környezetfüggetlen, hanem talán reguláris is egyben.

De térjünk vissza az "álmodozásaimból", és tekintsük át, mi lehet a közös a vizsgált jelenségekben? Milyen analógiák vonhatók a DNS-szekvenciák és a természetes nyelvi szövegek között, ami miatt hasonló jelenségeket fedeztünk fel bennük az előző fejezetben, valamint hasonló modellek alkalmazását javasoltam rájuk.

4. A genetikai kód és a természetes nyelvek közötti analógiák

Milyen alapon lehet egy kalap alá venni a DNS-szekvenciákat, a természetes nyelvi szövegeket, valamint − hozzá vehetjük még − a számítógépes programokat? Mindhárom "jelenséget" az a szerkezet jellemzi, amit az általam "*kettős kódolás*nak" ("*double coding*", a "kettős tagoltság" mintájára, ld. 1. fej.) nevezett folyamat hoz létre.

Vélemnyem szerint, a "*kettős kódolás*" lényege abból áll, hogy egy *megvalósítandó*t úgy kell átalakítani a *megvalósítóvá*, hogy egy *lineáris kód* formájában tároljuk az "információt". Linearitás alatt itt és a továbbiakban a kódot alkotó szimbólumok lineáris sorrendjét ertem, például azt, hogy a karaktereket az írás során egymás mellett helyezhetjük el. (Az írás történetének korai szakaszai mutatják, hogy ez a tény nem triviális.) A természetes nyelvi szövegek esetén a "megvalósítandó" a közölt információ (a mondat vagy a szöveg jelentése), míg a "megvalósító" a kiejtett hangsor. A DNS-szekvenciák esetén a "megvalósítandó"-t az enzim által ellátandó funkció jelenti, míg a "megvalósító" maga az enzim. Számítógépes program esetében pedig, a programozandó feladat a "megvalósítandó", míg a programkód a "megvalósító".

A "megvalósítandó" közös jellemzője mindhárom esetben a *nem-linearitás*. Egy mondat jelentése épp annyira komplex szerkezetű lehet, mint egy biológiai vagy számítógépes feladat. A "megvalósító" *kvázi-lineáris*. Ez azt jelenti, hogy a megvalósítót majdnem teljes mértékben leírhatjuk lineárisan. A "kvázi" jelzővel a fehérjék másodlagos és harmadlagos szerkezetére utalok, ill. arra, hogy a kiejtett hangkontinuumot igazából csak multilineárisan lehet leírni, a különböző hangképző szervek helyzete vagy egy hang fizikai tulajdonságai egymástól többé-kevésbé függetlenek. A "megvalósítandót" a "megvalósítóval" egy tökéletesen *lineáris kód* köti össze. A kérdés az, hogy miként lehetséges a nem-lineáris információt lineárissá átalakítani oly módon, hogy ne vesszen el semmi, azaz a megvalósítót a megvalósítandóhoz rendelő leképezés invertálható legyen. [1]

Ez a "kettős kódolás" legtisztábban a nyelvészet esetében figyelhető meg, ott összefüggésben van az emberi nyelv bevezetőben említett "kettős tagoltságával". A szemantika a nemlineáris jelentésből megalkotja a szintaxis nem-lineáris bemenetét, amit *kódolandó*nak fogok nevezni. A kódolandóból a szintaxis létrehozza a lineáris kódot. Ezt úgy teszi meg, hogy egy generatív grammatikát használ, mely szabályainak ismeretében, a kódból (a legenerált terminális-sztringből) visszafejthető a levezetési lánc (*parsing*), azaz a "mondat" nemlineáris szintaktikus struktúrája, ami viszont már a jelentés komplex szerkezetével függ össze. Például a

Béla fia szép fizikus lányt lát.

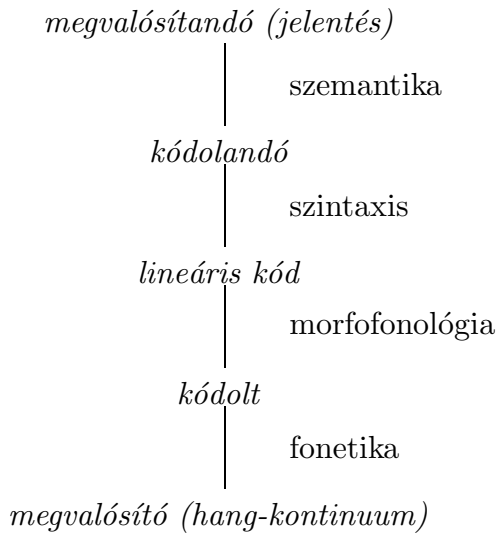
mondatból, a magyar nyelv szintaktikai szabályainak az ismeretében, rekonstruálható a mondatot alkotó elemek egymáshoz való viszonya. Utóbbit, zárójelek segítségével, az alábbi módon ábrázolhatjuk:

(Béla fia) ((szép (fizikus lányt)) lát).

Visszatérve "megvalósítandó" leképezésére "megvalósítóvá", a szintaxis kimenete a szavak (morfémák) lineáris sorrendje. Ez leképeződik hangalakokká, majd a morfológiai-fonológiai szabályok kisebb, lokális változtatásokat eszközölnek ezen a lineáris szekvencián (ún. *igazító szabályok*). A változtatások egy részének oka leginkább az emberi hangképző szervek tehetetlensége, például hasonulások esetén, vagy a percepció (megértés, rekonstruálás) elősegítése, például disszimilatív folyamatoknál. (A *természetes fonológia* terminusaival: szintagmatikai és paradigmatikai folyamatok [2] .) A hangsúly és a hanglejtés, és esetleg más artikulációs jegyek is (például az autoszegmentális fonológia megközelítésében), megtörik a linearitást, *egy* helyett *néhány* "tengelyre" lesz szükségünk. Ezért a fonológia kimenete, a *kódolt*, különösen pedig a fonetika immár nem is diszkrét elemekből álló, hanem folytonos kimenete, a *megvalósító*, multi- vagy kvázi-lineárisnak nevezhető. A 4.1 ábra foglalja össze ezt a folyamatot.

[1] Pontosabban szólva: a leképezés inverze *lehetőleg* ne legyen több értékű, és ha az esetek egy részében még több értékű is, akkor se legyen "sok" értékű. Hiszen a természetes nyelvekben előfordul a szerkezeti többértelműség. A "Péter és János kutyái" esetében nem tudjuk, hogy a "Péter + (János kutyái)" vagy a "(Péter és János) kutyái" jelentésre gondolt az, aki kimondta a szerkezetet. Hasonlóképpen, "Az oroszlán simogatása veszélyes" jelentheti egyszerre azt, ha az oroszlán simogat, és azt is, ha az oroszlánt simogatják.

[2] Ld. Kiefer [1994], p. 39.



4.1 ábra: A kettős kódolás szintjei

Fontos megjegyezni, hogy a fenti ábra mindkét irányba járható, hiszen a beszédértés éppen a beszédprodukcióval ellentétes folyamat.

A genetikában még nem ismerjük kellőképpen a részleteket. A "lineáris kód" szerepét nyilvánvalóan a DNS-szekvencia tölti be, ez kódolja a genetikai információt, hogy az a létrejövő fehérjék elsődleges, másodlagos és harmadlagos szerkezete révén valósuljanak meg. A genetikai információ, mint "megvalósítandó", nyilván meglehetősen komplex. Ebben a képben a sejtbiológia felel meg a szemantikának, hiszen a sejtben lezajló folyamatok, a "kódolandók", jelentik az első (utolsó) lépést az információ formába öntése felé. A genetika (a biológiai rendszerek szintaxisa) átkódolja a "folyamatokat" DNS- szekvenciává. A fehérje-szintézis előbb lokális változtatásokat hajt végre ("fonológia"), például kihagyja az intronokat [3], az m-RNS tekinthető "kódoltnak", majd a riboszómákon "materializálódik" az információ, ez az utolsó szakasz ("fonetika") hozza létre a megvalósítandó genetikai információt *megvalósító* enzimet. Az enzimológia bezárja a kört, mivel megadja azt, hogy hogyan valósítja meg a megvalósító a megvalósítandót. Ez a fázis a természetes nyelvek esetén hiányzik, egyedül a kettős tagolást nem tartalmazó kommunikáció (pl. indulatszavak, állati kommunikáció, gesztusnyelv,...) létezik szoros kapcsolat a hang és a jelentés között.

(A fenti gondolatmenetben meglepő lehetett, hogy a genetikát mint a DNS-kód tudományát említettem, és például a fehérjeszintézis kutatását kizártam a genetikából. Lehet, hogy ez nagyon merész lépés, és jobb terminusokat kellett volna találnom, mint például "genetikai szintaxis". De a döntésem oka a nyelvészeti analógia volt, ahol chomskyánus felfogás a szintaxist tekinti a nyelvészet magjának. A másik oka az volt, hogy feltételezem az egész dolgozatom során azt, hogy a DNS "nem-kódoló" részei valójában nem "evolúciós

[3] Lehet, hogy ezen transzformációk a transzformációs grammatikák analogonjai? A felvetés valószínűleg túlságosan messze vezetne, és a mai genetikai ismereteink nem teszik lehetővé, hogy megalapozott választ adjunk erre a kérdésre.

szemét", hanem van szerepe, és olyan információt kódol, ami túlmegy a fehérjék elsődleges szerkezetét kódoló, ismert mechanizmusokon. Amennyiben tényleg így lenne, úgy a jövő genetikai kutatásai a DNS-kód a jelenleginél mélyebb megértéseit fogják eredményezni.)

Számítógépes programok esetén, a gondolatmenet hasonló a fentiekhez. A feladat (a megvalósítandó) megfogalmazása (kódolandó) után, a programozás folyamata nem más, mint egy lineáris kód létrehozatala, a számítógépes nyelv szintaxisa segítségével. A fonológiának talán az algoritmus leíró nyelv elvont utasításainak kódolása jelenti, a programozási nyelv konkrét karaktersorozataként. Lokális változtatásokat jelent például megjegyzések beszúrása. Ha az algoritmus leíró nyelv felel meg a Chomsky-féle univerzális grammatikának, akkor megfigyelhető, hogy a szintaxis kisebb, és a "fonológia" nagyobb része nyelvspecifikus, mind a természetes, mind a számítógépes nyelvek esetén. Ez a meglehetősen általánosító megállapítás érthető, ha arra gondolunk, hogy a "megvalósítandó" még teljesen független az alkalmazandó nyelvtől, míg a "megvalósítót" a nyelv hozza létre. Minél közelebb vagyunk a "megvalósítóhoz", annál több nyelvspecifikus jelenségre számíthatunk, mivel a "megvalósítás" részeredménye az adott szinten maga is annál inkább nyelvspecifikus.

Hogyan lehet a fenti gondolatmenetet igazolni, honnan tudjuk, hogy ilyen módon magyarázhatóak a DNS-szekvenciák és írott nyelvi szövegek elemzésénél talált hasonló jelenségek? (Programozási nyelvekkel a továbbiakban nem foglalkozom.)

A *megvalósítandó* átfogalmazása *kódolandó*vá egy logikai struktúra létrehozását jelenti. Ez feleltethető meg a szintaxis által használt környezetfüggetlen generatív grammatika levezetési fáinak, gondoljunk csak a fenti, zárójelezett példamondatra. (A levezetési fák ekvivalensek a zárójelezéssel.) Ezzel a megfeleltetéssel a dolgozatomban nem foglalkozom részletesen. Tehát az elemi összetevők közötti komplex struktúrához egy-az-egyben rendelhető levezetési fa (parse-tree), melyhez pedig egy-az-egyben (esetleg "néhány-az-egyben") rendelhető egy lineáris sorrend, *adott (rögzített) grammatika mellett.* Vagyis a generatív grammatika teszi lehetővé a nem-lineáris szerkezet kódolását lineáris kód formájában. Ez a generatív grammatika viszont nem lehet reguláris, hiszen a komplex szerkezet távoli összetevők közötti kapcsolatot is lehetővé tesz, amelyet véges sok állapottal nem lehet elemezni. A gondolatmenetet folytatva, ennek a következménye az, hogy a szintaxist nem lehet reguláris grammatikával leírni, legalább környezetfüggetlen nyelvtan kell hozzá. és az irregularitás, azaz a memória-igény felbukkanásának következménye a hosszútávú korrelációk megjelenése a szövegben. (A fenti gondolatmenet matematikai megfogalmazása hiányzik, de remélem, hogy megtehető.) E ponton kapcsolható a dolgozatom a statisztikus fizikához, hiszen a fizika nyelvén szólva, komplex struktúrák hosszútávú korrelációt létrehozó mechanizmusát vizsgálom.

A lineáris kód kódolttá és megvalósítóvá történő transzponálása során viszont csak rövidtávú, lokális ("kényelmi") változásokat hajtunk végre nyelvek esetén. A genetika nem tud még választ adni arra a kérdésre, hogy miért van szükség ezekre a változtatásokra (nem-kódoló részek kihagyása,...). A programozási nyelvek esetén nincs is szükség ilyen változtatásokra, hiszen − mesterségesen létrehozott nyelvről lévén szó − a nyelvet létrehozó személy általában nem definiált ilyet.

Ezek a lokális változtatások leírhatók reguláris nyelvvel. Igaz, hogy a szokásos fonológiai formalizmus látszólag nemcsak, hogy nem reguláris, hanem nem is környezetfüggetlen, hanem környezetfüggő, hiszen egy tipikus klasszikus generatív fonológiai szabály alakja:

$$A \to B \setminus C\_D,$$

amely a formális nyelveknél megszokott alakban felírt

$$CAD \to CBD,$$

környezetfüggő szabálynak felel meg. Mégis, a fonológiai szabályokra, ill. azok működésére általában olyan megszorításokat szoktak alkalmazni, amelyek regulárissá teszik az általuk leírt nyelvet (Kaplan & Kay [1994]).

A következő fejezetben belátjuk, hogy a reguláris grammatikák ekvivalensek a Markov-modellekkel, vagyis nem eredményezhetnek hosszútávú korrelációkat. Ezek után érthetővé válik, hogy miért veszítettük el a hosszútávú korrelációkat a szövegek megpermutálá- sakor: a permutáció tönkreteszi szintaxis a nem reguláris nyelvét, és csak a szó szintű, rövid távú fonológiai korrelációk maradtak meg.

Mi történik, amikor c-DNS-t készítünk? Szintén a "genetikai fonológia", azaz a transkripció eredményét vizsgáljuk, és az ezen a szinten működő szabályok már nem hoznak létre hosszútávú korrelációkat. Arra viszont nem tudok választ adni, hogy vajon miért szűnnek meg itt a korábban létrejött hosszútávú korrelációk, holott azok megmaradnak a permutálatlan hangsorban. Valószínűleg a választ csak a DNS-kód jobb megértése, a "nem-kódoló" részek szerepének tisztázása adja meg.

5. Sztochasztikus formális nyelvek és Markov-modellek

A generatív nyelvészet által használt formális nyelvek elméletének fentebb ismertetett formája nem teszi lehetővé, hogy a korrelációk statisztikus jelenségének nyelvészeti vonatkozásait megvizsgálhassuk, vagy a korrelációk létére nyelvészeti magyarázatot találjunk. Ehhez az szükséges, hogy a formális nyelvek elméletét kiegészítsük sztochasztikus eszközökkel. Először a *sztochasztikus környezetfüggetlen grammatikák* néven ismert modellt ismertetem, Krenn & Samuelsson [1996] alapján. Majd ennek speciális esetét, a sztochasztikus reguláris grammatikákat vizsgálom meg, és megmutatom kapcsolatukat a Markov-modellekhez.

5.1 Sztochasztikus környezetfüggetlen grammatikák

Ha a generatív grammatikák négyesét egy valószínűségi függvénnyel egészítjük ki, sztochasztikus grammatikákat kapunk. Ilyen módon nem csak azt mondhatjuk meg, hogy egy adott mondat vajon eleme-e a grammatika által generált nyelvnek, hanem azt is, hogy milyen valószínűséggel vezethetjük le az S mondatszimbólumból az adott szekvenciát. Ezt a valószínűséget úgy értem, hogy amennyiben összefűzzük a "végtelen" sok mondatát, azaz képezünk egy *ideális szövegkorpuszt*, akkor az adott mondat milyen gyakorisággal fordul elő a korpuszban. Az ideális korpusz jó közelítését adhatja természetes nyelvek esetén egy könyv (pl. egy regény), a "DNS-nyelv" esetén pedig a DNS-szekvenciákat tartalmazó adatbázisokat fogom ilyen korpusznak tekinteni.

(Természetes, hogy a választott korpusz befolyásolhatja a valószínűségeket. Például az "*A Maxwell-egyenletek nem invariánsak a Galilei-transzformációra.*" mondat valószínűsége

nyilvánvalóan más a Fizikus Tanszékcsoport könyvtári anyagából képezett korpuszban, mint a Csepeli Fémművek dolgozói által 1996. során kiejtett mondatok, mint korpusz esetében. De ez nem zavar bennünket, hiszen a nyelvészek gyakran kényszerülnek leírásaikat egy szűkebb körre, például a magyar nyelv esetében a "budapesti köznyelvre", leszűkíteni. A valószínűségek egy részének korpusz-választástól való függését ezért szociolingvisztikai kérdésnek tekinthetjük. Ennél nehezebb kérdés az, hogy miként definiálhatunk egy közel ideális korpuszt. Nyilván ilyen lenne a budapesti köznyelv esetén a következő meghatározás: "az 1996. december 31-én 18. életévüket betöltött, érettségizett, magyar anyanyelvű budapesti lakosok által 1997. során kiejtett mondatok." Viszont ezen korpusz kísérletileg nem vizsgálható. Amennyiben viszont az 1980-as években megjelent sajtótermékek korpuszát tekintjük, az anyag inkább az irodalmi, mintsem a köznyelvi állapotokat fogják tükrözni. Természetesen, ennek a vizsgálata is érdekes lehet.)

5. Definíció: Sztochasztikus környezetfüggetlen grammatikának (Stochastic Context-free Grammar, SCFG) nevezzük az (V_N, V_T, S, R, P) *ötöst, ahol:*

V_N *a nemterminálisok véges halmaza,*

V_T *a terminálisok véges halmaza (legyen újból* $V := V_N \cup V_T$*),*

$S \in V_N$ *a nemterminálisok közül a kitüntetett mondatszimbólum,*

R *az* $X \to Q$ *alakú levezetési szabályok egy véges halmaza, ahol* $X \in V_N$ *és* $Q \in V^*$*,*

P *pedig egy* $R \to [0,1]$ *függvény oly módon, hogy*

$$\forall X \in V_N : \sum_{Q \in V^*} P(X \to Q) = 1.$$

A $P(X \to Q)$ valószínűség azt jelenti, hogy ha egy levezetési láncban megjelenik egy X nemterminális, azt milyen valószínűséggel fogjuk Q-ként újraírni.

Ezek után, egy levezetési lánc, ill. levezetési fa valószínűségét úgy definiálhatjuk, mint az alkalmazott helyettesítési (újraíró) szabályokhoz rendelt valószínűségek szorzatát. Valamely mondat valószínűsége pedig a hozzátartozó levezetési fák (*parse trees*) valószínűségeinek az összege lesz. A módszer a számítógépes nyelvészetben hasznos elemző automaták (*parser*-ek) futási idejének optimalizálására. Ha a korpuszt úgy tekintjük, mint nagy számú, egymás mellé írt mondatszimbólumból ("poli-S láncból") levezetett sztring, úgy látható a kapcsolat a korpuszból számított empírikus gyakoriság és a fentebb bevezetett elméleti valószínűség között.

Mivel a DNS-szekvenciákkal és természetes nyelveken írt szövegekkel foglalkozó fizikusok egyik régi érdeklődési területe a *Zipf-törvény*:

Zipf-törvény: Ha összeszámoljuk egy adott szövegben a különböző szavak előfordulási gyakoriságát, majd a csökkenő gyakoriság szerint sorba rendezzük őket, akkor a sorrendben elfoglalt R *hely függvényében ábrázolva az* $\omega(R)$ *gyakoriságot (frekvenciát), egy, közelítőleg* -1 *exponenssel jellemezhető hatványfüggvényt kapunk.*

(Zipf, [1935, 1949], Czirók[1995, 1996], Kanter & Kessler [1994])

én az egyes terminálisok előfordulási valószínűségére voltam kíváncsi egy korpuszban, azaz mondatok láncában. Ezért végeztem el a következő számolást: $v_S(a)$-val jelölöm azt, hogy

az $a \in V_T$ terminális várhatóan hányszor fordul elő egy S-ből levezethető mondatban [1]:

$$v_S(a) := \sum_{\sigma \in V_T^*} p_S(\sigma) \binom{\sigma}{a},$$

ahol $p_S(\sigma)$ jelöli a $\sigma \in V_T^*$ mondat fentebb bevezetett valószínűségét, $\binom{\sigma}{a}$ pedig az a terminálisok számát a σ mondatban. (Megjegyzem, hogy habár formálisan nem látom be, de ezen végtelen szummának konvergensnek kell lennie. Hiszen felülről becsülhető a mondatok hosszának várható értékével, amely viszont véges, hiszen enélkül a kommunikáció nem lenne elképzelhető.)

A bennünket érdeklő $v_S(a)$ várható értéknél többet fogunk kiszámítani a SCFG paramétereiből, azaz a P valószínűségfüggvényből. Jelölje $p_A(\sigma)$ bármely $A \in V_N$ nemterminálisra annak a valószínűségét, hogy A-ból a $\sigma \in V_T^*$ sztringet vezetjük le: az A gyökerű, σ-t eredményező levezetési fák valószínűségeinek footnote[2]Az ilyen fák valószínűségét szintén az alkalmazott levezetési szabályok valószínűségeinek a szorzataként számíthatjuk ki, akárcsak az S gyökerű fák esetén. az összegét. és jelölje $v_A(a)$ annak a várható értékét, hogy az A-ból sztringekben hányszor szerepel az a terminális:

$$v_A(a) := \sum_{\sigma \in V_T^*} p_A(\sigma) \binom{\sigma}{a}. \tag{5.1}$$

Ezek a $v_A(a)$-k egy $\mathbf{v}(a)$ vektornak a komponensei, amely egy, a V_N elemeinek a számával megegyező dimenziójú térben van.

A $p_a(\sigma)$ átírásához be kell vezetni a *Chomsky-féle normálalak* fogalmát. Belátható (pl. ld. Révész [1979], SCFG-re Krenn & Samuelsson [1996]), hogy bármely környezetfüggetlen grammatikához létezik olyan Chomsky-féle normálalakban megadott grammatika (*Chomsky Normal Form*, *CNF*), amely ugyanazt a nyelvet generálja, és amelynek a szabályai

$$A \to BC$$

vagy

$$A \to a$$

alakúak, ahol $A, B, C \in V_N$ és $a \in V_T$. Tetszőleges SCFG-hez is adható meg olyan SCFG, amelynek R-je CNF-alakú levezetési szabályokat tartalmaz, és amely minden sztringet ugyanolyan valószínűséggel generál, mint az eredeti SCFG.

Definiáljuk még a $\pi(a)$ vektort és a μ mátrixot:

$$\pi_A(a) := P(A \to a)$$
$$\mu_{AB} := \sum_{C \in V_N} \Big[P(A \to BC) + P(A \to CB) \Big]$$

[1] Az alábbiakban a kisbetűk mindig terminálisra, a nagybetűk nemterminálisra, míg a görög betűk terminálisokból álló sztringre fognak utalni.

Amikor az A-ból akarom levezetni a σ-t egy CNF alakú grammatikában, két lehetőségem van az elindulásra: amennyiben a σ egyetlen a terminálisból áll, úgy az $A \to a$ szabályt kell alkalmaznom, egyébként pedig az A-t egy $A \to BC$ szabállyal írom újra, és B-ből levezetem σ_1-et, C-ből σ_2-t, ahol $\sigma = \sigma_1\sigma_2$ alakú (e kettő konkatenációja). Ennek megfelelően:

$$p_A(\sigma) = \sum_{b \in V_T} P(A \to a)\delta_{\sigma,b} + \sum_{B,C \in V_N} \sum_{\substack{\sigma_1,\sigma_2 \in V_T^* \\ \sigma = \sigma_1\sigma_2}} P(A \to BC)p_B(\sigma_1)p_C(\sigma_2) \tag{5.2}$$

(A $\delta_{\sigma,b}$ szimbólum a megszokott Kronecker-deltát jelöli.) Ha beírjuk a (5.2) egyenletet (5.1)-be, úgy bizonyos egyszerűsítésekre lesz lehetőségünk, a következő három összefüggés felhasználásával:

$$\left(\frac{b}{a}\right) = \delta_{b,a}$$

$$\sum_{\sigma \in V_T^*} p_C(\sigma) = 1$$

$$\left(\frac{\sigma}{a}\right) = \left(\frac{\sigma_1}{a}\right) + \left(\frac{\sigma_2}{a}\right)$$

Az eredmény a következő lesz:

$$v_A(a) = \pi_A(a) + \sum_{B,C \in V_N} P(A \to BC)(v_B(a) + v_C(a)),$$

amelyből:

$$v_A(a) = \pi_A(a) + \sum_{B \in V_N} v_B(a)\mu_{AB},$$

vagyis:

$$\mathbf{v}(a) = \pi(a) + \mu\mathbf{v}(a). \tag{5.3}$$

A (5.3) összefüggés átrendezésével a SCFG paramétereiből ki tudjuk számítani a Zipf-törvényben szereplő frekvenciákat. Ehhez az szükséges, hogy az $1 - \mu$ mátrix invertálható legyen, de ezzel azért nincs baj, mert a μ elemeire nincsen semmiféle megkötés ($\pi(\mathbf{a})$ kis megváltoztatására μ is megváltozik), vagyis kis perturbációval megszüntethető egy esetleges szingularitás.

5.2 Sztochasztikus reguláris nyelvtanok és a Markov-modell

Az előző részben bevezetett SCFG speciális esete az, ha nem csupán a környezetfüggetlenséget, hanem a regularitást is kikötjük. Jelen fejezetben az 5. definíció szerinti P valószínűséggel kibővített *sztochasztikus reguláris nyelvtan*okra (*SRG*) fogom belátni, hogy helyettesíthetők

Markov-modellekkel. (A Markov-folyamat fogalmát nem definiálom, mivel a fizikusok "műveltségébe" beletartozik. Jó és tömör leírása megtalálható Krenn & Samuelsson [1996]-ban.)

Előljáróban megjegyzem, hogy egy *reguláris korpusz*, pontosabban szólva: egy reguláris nyelv mondataiból összefűzött szöveg maga is tekinthető egyetlen reguláris mondatnak. [2] Elegendő, ha az eredeti generatív grammatika minden $A \rightarrow a$ alakú, mondatgenerálást lezáró szabálya mellé felveszünk egy $A \rightarrow aS$ szabályt is, mely az előző mondatot lezárja, és egy újat kezd el, a szöveg következő mondatát. Tehát, ha egy mondat-lánc regularitását vizsgáljuk, nyugodtan tekinthetjük az egészet egyetlen mondatnak. (Olyan ez, mintha pontok helyére pontosvesszőket tennénk.)

Először, értsük meg, miben különbözik egy Markov-modell egy SRG-tól, illetve az azzal ekvivalens véges állapotú automatától, amennyiben az utóbbit szintén felruházzuk átmeneti valószínűségekkel. Mindkettő egy véges állapothalmazból vett állapotok sorozatán halad keresztül, két állapot közötti átmenet során kibocsát egy-egy jelet, és az átmenetek, ill. a jelkibocsátás bizonyos valószínűséggel következik be.

Viszont, míg a Markov-modell esetén független a jelkibocsátás az átmenettől, addig a SRG esetében a kettő *együtt* következik be: $P(A \rightarrow aB)$ annak a valószínűsége, hogy az automata az A állapotból a B állapotba megy át, *miközben* az a jelet bocsátja ki. Más szavakkal: az SRG "élcimkézett" (*arc-labeled*), míg a Markov-modell "állapotcimkézett" (*state-labeled*), vagyis a SRG-nál (azaz a véges állapotú automatánál, *finite state automaton*, *FSA*) a kibocsátott jel (a "cimke") az állapotok közötti élhez kapcsolódik, míg a Markov-modelleknél magához az állapotokhoz. Viszont tudjuk, hogy a "state-labeled" és az "arc-labeled" automaták ekvivalensek (ld. pl. Bird & Ellison [1994]). Ennek felel meg a következő állítás a sztochasztikus automatákra (nyelvekre, modellekre) [3]:

Állítás: Minden sztochasztikus reguláris nyelv megvalósítható Markov-modellel.

A bizonyítás során megkonstruáljuk a megfelelő Markov-modellt. A Markov-modell állapotai között egyrészt megtaláljuk a SRG nemterminálisainak (a véges állapotú automata állapotainak) megfelelő állapotokat, másrészt a SRG minden helyettesítési szabályának is megfelel egy állapot. Előbbieket nevezzük *primer*, utóbbiakat pedig *szekunder* állapotnak. A Markov-modell kimenő ábécéje természetes módon megegyezik az SRG terminálisaival (a leírandó nyelv ábécéjével), illetve tartalmaz még egy λ elemet is, melynek jelölésére nem véletlenül választottam a lambdát, hiszen sokan ezt használják az üres szó jelölésére (amire mi az e-t választottuk): a λ szerepe az lesz, hogy ignoráljuk őt.

Meg kell még adnunk a Markov-modell átmeneti valószínűségeit és a jel-kibocsátási valószínűségeit. Primer állapotból csak szekunder állapotba, szekunder állapotból csak primerbe mehetünk át nem-zérus valószínűséggel. Az A nemterminálisnak megfelelő állapotból a $B \rightarrow aC$ állapotba való átmenet valószínűsége $\delta_{A,B}P(B \rightarrow aC)$. A $B \rightarrow aC$ állapotból az A állapotba pedig: $\delta_{C,A}$. Primer állapotban 1 valószínűséggel bocsátja ki a Markov-modell a λ jelet (más olvasat szerint: nem bocsát ki jelet), míg minden mást zérus valószínűséggel. Amikor a Markov-modell által generált jelsorozatot vizsgáljuk, egyszerűen figyelmen kívül

[2] Azaz, ha L reguláris nyelv, akkor L^* is reguláris, ld. Révész [1979], *2.2 tétel*.

[3] A Markov-modell lényegében egy visszafele működő, sztochasztikus eszközökkel kibővített *state-labeled automaton*.

hagyjuk a páratlan helyeken megjelenő λ-kat. Az igazi jelet a szekunder állapotokban bocsátja ki, a $B \to aC$ szabálynak megfelelő állapotban egyedül az a jelet bocsátja ki nem-zérus valószínűséggel (1 valószínűséggel).

Könnyen látható, hogy az így vázolt Markov-modell ekvivalens az eredeti SRG-val, ha eltekintünk a λ-k generálásától, valamint a SRG-t átalakítjuk a fejezet elején említett módon mondatsorozat (szöveg) generálására, és "végtelenítjük" a működését. (Nem használjuk az $A \to a$ alakú szabályokat; ill. a használatuk a Markov-modellt egy végállapotba viszi, ahol megáll a működése.) A Markov-modell két lépésben teszi meg ugyan azt, és ugyan olyan valószínűséggel, mint amit a SRG egy generálási lépés alatt, illetve az SRG-vel ekvivalens véges állapotú automata egy lépés alatt.

Ezek után összefoglalhatjuk eddigi eredményeinket: a formális nyelvek elmélete alkalmas egy sor jelenségkör leírására (természetes nyelvek, programozási nyelvek, valószínűleg a DNS és a fehérjék szerkezetének leírására is). A formális nyelvek elméletét kibővítve sztochasztikus módszerekkel, azt kaptuk, hogy a reguláris grammatikák ekvivalensek a Markov-modellekkel. Mivel a Markov-láncok, és így a Markov-modellek sem, képesek hosszútávú korrelációk produkálására, bizonyítottnak vehetjük, hogy azok a jelenségek, amelyeknél hosszútávú korrelációkat észleltek (intront tartalmazó DNS-ek, természetes nyelvi szövegek a szó szintje fölött,...) nem elemzhetőek reguláris grammatikákkal. Ez a szintaxis esetén nem újdonság, hiszen már Chomsky [1957] is emellett érvel. A genetikában bármely elméletnek, mely az intronok szerepére vonatkozik, összhangban kell lennie ezzel az elmélettel (Mantegna et al. [1995a]). Azok a szintek, amelyek nem mutattak hosszútávú korrelációkat (fonológia, c-DNS-ek,...) viszont esetleg leírhatók reguláris eszközökkel, mint ahogy a fonológia esetében történtek is erre már kísérletek.

A következő két fejezetben bemutatandó "vektortér-technika" − M. Damashek [1995], az alapötlet kidolgozója nevezte −, a szimbólumsorozatok reguláris viselkedését veszi közelebbről szemügyre.

6. n-grammok és fonotaktika

Az elmúlt években egyre több algoritmus születik szövegek automatikus szelektálására: például egy hírügynökség számára nagyon hasznos lenne, ha a befutó híreket emberi beavatkozás nélkül lehetne nyelv és témakör szerint csoportosítani. Damashek [1995] éppen egy ilyen algoritmusra tesz javaslatot, melyet ő *Acquaintance*nek nevez.

Ennek az algoritmusnak a lényege az, hogy a szövegekből vektort készít, majd a vektorok skaláris szorzata jellemző a vektorok, azaz a szövegek hasonlóságára.

Képzeljük el, hogy egy n hosszúságú ablakot ($n = 3..6$) mozgatunk végig a szövegen, karakterről karakterre. A különböző lehetséges karakter-n-eseket indexeljük $i = 1..J$-vel. Jelöljük m_i-vel az i-ik karakter-n-es előfordulásainak a számát. A szövegből képezett $\mathbf{x}$ vektor i-ik komponensét éppen az m_i lenormálásából kapott frekvenciaként képezzük:

$$x_i := \frac{m_i}{\sum_{j=1}^{J} m_j} \tag{6.1}$$

Damashek praktikus tanácsot is ad ezen vektor gyors és memóriatakarékos kiszámítására: elégséges csupán a zérustól különböző vektorkomponenseket tárolni; a vektort pedig úgy állíthatjuk elő, hogy a karakterszekvenciából képezzük az n-esek szekvenciáját, majd egy hatékony rendezési algoritmussal rendezzük ezt a szekvenciát, és ezt követően elég összeszámolni, mely n-esből mennyi van a szövegben.

A két szövegből ilyen módon képezhető vektorok, $\mathbf{x}$ és $\mathbf{y}$, skaláris szorzata jellemzi a szövegek hasonlóságát:

$$S = \frac{\sum_{j=1}^{J} x_j y_j}{\left(\sum_{j=1}^{J} x_j^2 \sum_{j=1}^{J} y_j^2\right)^{1/2}} = \cos\theta \tag{6.2}$$

Könnyen belátható, hogy a $d(\mathbf{x}, \mathbf{y}) := 1 - S$ távolság egy metrikát definiál a vektorok terében, azaz jó távolságfogalom a szövegek között. (Attól a valószínűtlen esettől eltekintve, amikor két különböző szöveg ugyanarra a vektorra képezhető le.)

Ez a szorzat elsősorban a szövegek nyelv szerinti szétválasztására alkalmas, de szerencsés esetben az azonos nyelvű szövegek téma szerinti szétválasztása is lehetséges szerencsés esetben. Saját magam például öt, fizikai témája angol e-mailből (Ph1-Ph5), három szintén angol nyelvű, de más témájú e-mailből (E1-E3), két francia levélből (Fr1-Fr2) és három magyar nyelvű, magánjellegű e-mailből (H1-H3) álló korpuszt vizsgáltam. Azt egyes szövegek hossza 3400 és 6000 karakter között van, kivéve a két francia levelet, amelynek hossza csupán 1000 - 1200 karakter. Az angol ábécé 26 betűjét, a pontot, a vesszőt, valamint az üres helyet vettem figyelembe. A többi karaktert ignoráltam, kis- és nagybetű között, ill. ékezetes és ékezet nélküli betű között pedig nem tettem különbséget. A több üres helyből álló szekvenciákat előzőleg törölni kell. Eredményeimet $n = 3$-ra és $n = 4$-re az 6.1, ill. a 6.2 táblázat tartalmazza.

(Damashek a módszert továbbfejleszti. Bevezet ún. *centroid vektorok*t, amely egy szövegcsokor (például az angol nyelvű szövegek, vagy az angol nyelvű, számítástechnikai témájú szövegek) vektorainak az átlaga. Ezt a vektort levonva a szövegek vektoraiból, kitranszformálhatóak a szövegcsokor közös jellemzői, például az adott nyelv funkcionális-grammatikai szavaiból adódó jellegzetességek (a magyar esetén pl. az " a " hármas vagy az " az " négyes). Ilyen módon, a szövegek finomabb csoportosítása is lehetséges. Damashek javasol praktikus ötleteket a statisztikus hibák kikerülésére is.)

A két táblázat adatai között szignifikánsan magasabbak az azonos nyelvű szövegek vektoraiból képezett vektorok szorzatai, mint a különböző nyelvűek szorzatai. A két, különböző témájú angol szövegcsokor is elkülönül egymástól: a Ph-szövegek (pl. $n = 3$-ra a Ph-szövegek egymás közötti szorzata: $0,79 \pm 0,024$) ill. az E-szövegek szorzata $(0,80 \pm 0,015)$ magasabb, mint egy E- és egy Ph-szöveg szorzata$(0,68 \pm 0,03)$.

Ezt követően, összefűztem mind a négy szövegtípusból néhány szöveget, egyetlen fűzérré. [1] Egy $m = 100$ hosszúságú dobozt mozgattam végig ezen a fűzéren, és a doboz

[1] A file szerkezete: 0-6775. karakter: Ph-típus; 6776-13551. karakter: E-típus; 13552-23219. karakter: Ph-típus; 23220-25862. karakter: H-típus; 25863-32319. karakter: Ph-típus; 32320-35178. karakter: Fr-típus.

éppen aktuális tartalmából képezett vektort ($n = 3$) összeszoroztam a fűzér elejéből (0-6775. karakterek közötti Ph-szövegekből) képezett vektorral. Az eredmények az 6.1. ábrán láthatóak: a fűzér angol és nem angol nyelvű részei élesen különválnak.

Az eredmény magyarázata ott keresendő, hogy különböző nyelvekre különböző karakterszekvenciák jellemzőek. Például az angol nyelvű szövegekben, az $n = 3$ esetben kiugrik a "the" karakterhármas: gondoljunk a határozott névelőn kívül a névmásokra ("they", "their", "them", "these", "there", stb.).

Véleményem szerint, ez a jelenség részben ortográfiai okora megy vissza (pl. az "sz" pár csak azért jellemző a magyar szövegekre, mert a magyar helyesírás így jelöli a fonetikai [s] hangot). Másrészt, igazi nyelvészeti okokat is találhatunk, hiszen az írott szöveg részben tükrözi a kiejtett szöveg fonológiai jellemzőit.

A Damashek-féle vektorok jellegzetességei részben a nyelv *fonotaktikájá*ra mennek vissza. A *fonotaktika* a fonológia azon ága, amely a fonémák lehetséges kapcsolatait vizsgálja a célnyelvben (Kiefer, [1994], 4. fejezet). Például az **bárdás* szó nem létezik a magyar nyelvben, de nem érezzük idegennek, "akár létezhetne is", hiszen megfelel a magyar nyelv fonotaktikai szabályainak. Szemben az előforduló, de nem jólformált *görl* (pl. show-girlök) szóval.

A fonotaktika jellegzetesen "reguláris" vizsgálódási terület. A fonotaktikai megszorítások felírhatók reguláris grammatikák formájában, melyek megengedik a jólformált alakokat, és kizárják a nem jólformáltakat.

A fonotaktikai elemzések során gyakran alkalmaznak a nyelvészek kisebb "csalásokat". Előforduló alakokat néha nem-jólformáltnak tekintenek: ritkán előforduló, idegen hangzású szavak (pl. **ng***anaszán* [szamojéd nyelv neve], **scs***í*, *fá***jl**) esetén még ki lehet magyarázni azzal, hogy felejtsük el őket. De a *bara***ck**, *rece***pt**, *tere***mt** szavak esetén már nincs ilyen egyszerű dolgunk.

A "csalások" másik része az, amikor nemlétező hangkapcsolatokat jólformáltnak tekintünk az elemzés egyszerűsítése miatt, és csupán "véletlen hiánynak" tekintjük azt, hogy például a magyarban nincsen [t^{y}a]-val kezdődő szó. Ezen hiányokat "érzés" alapján tekintjük helyesnek, de mit jelent az, hogy "véletlen hiány"?

Ezért javaslom a fonotaktika sztochasztikus megközelítését: egészítsük ki a fonotaktika reguláris grammatikáját sztochasztikus eszközökkel, azaz térjünk át Markov-modellre. Ilyen megközelítésben, jólformált az az alak, amelynek nagy valószínűséget jósol a modell, a nem-jólformált alakokhoz pedig elhanyagolható (zérus) valószínűséget rendelünk. A "csalások" magyarázhatóak az elméleti valószínűség körüli statisztikus szórással. Az agrammatikus, előforduló alak olyan szekvenciának felel meg, amely az elmélet által jósolt valószínűségnél nagyobb gyakorisággal fordul elő, míg a véletlen hiányok a számítottnál alacsonyabb gyakoriságot jelent.

A fonotaktika sztochasztikus megközelítése jólformáltsági fokozatok megkülönböztetését teszi lehetővé. Például a [ck#] szóvéget kevésbé érezzük nem-jólformáltnak, mint a [#ng] szókezdetet. Az előbbi több, és gyakoribb szavakban fordul elő (*barack*, *palack*, *tarack*,...),

mint az utóbbi (*nganaszán*). A sztochasztikus modellek ezen fokozatok leírását lehetővé teszik.

Egy megjegyzés a gyakoriságok meghatározásáról. A nyelvészek általában "lexicon-based" statisztikákat használnak, mivel őket csak az érdekli, vajon egy alak előfordul-e vagy sem (ld. pl. a Kiefer [1994] 4. fejezetében használt 80 000 tételből álló adatbázist). Ezzel szemben, a Damashek-féle módszer ún. "corpus-based" eljárás, hiszen a szövegben gyakrabban előforduló szavak, hangkapcsolatok nagyobb súllyal szerepelnek. Ennek előnye az, hogy különbséget tud tenni például a [#ng] szókezdet és a [mt#] szóvég között: habár mindkettőt agrammatikusnak tartják az elemzések, mivel mindkettő csupán egyetlen mag-

yar szóban fordul elő (*nganaszán*, ill. *teremt*), de az utóbbi sokkal gyakrabban fordul elő a korpuszokban, és az anyanyelvi beszélő és érzi a két alak közötti jólformáltsági fokozatot.

Tegyük fel, hogy a nyelvünk fonotaktikáját le tudjuk írni egy olyan Markov-modellel, amely N állapotot $(s_1, s_2, ..., s_N)$ és M darab jelet $(\sigma_1, \sigma_2, ..., \sigma_M)$ tartalmaz. Az i-ikből a j-ik állapotba történő átmenet valószínűségét jelöljük p_{ij}-vel, míg a_ij annak a valószínűsége, hogy az i-ik állapotban a modell a j-ik jelet bocsátja ki. Tegyük fel, hogy ergodikus a Markov-modellünk; és mivel hosszú láncokról, reguláris nyelv hosszú modatairól van szó (ne felejtsük el, hogy a reguláris nyelvek osztálya zárt a * műveletre, ld. 5.2 fejezet elején), feltehetjük azt is, hogy a modell már "egyensúlyivá vált", azaz annak a v_i valószínűsége, hogy a szekvencia valamely pozíciójában a rendszer az i-ik állapotban lesz, független a pozíciótól. Ekkor a $\sigma_{k_1}\sigma_{k_2}...\sigma_{k_n}$ szimbólum n-es elméletileg jósolt valószínűsége:

$$P(\sigma_{k_1}\sigma_{k_2}...\sigma_{k_n}) = \sum_{i_1=1}^{N} v_{i_1} a_{i_1 k_1} \sum_{i_2=1}^{N} p_{i_1 i_2} a_{i_2 k_2} ... \sum_{i_1=1}^{N} p_{i_{n-1} i_n} a_{i_n k_n}$$

Mivel a p_{ij} és a_{ij} paramétereket nem ismerjük, a jövőbeni kutatások érdekes feladata lehet valamely adott korpusz esetén ezen "rejtett Markov-modell" (*hidden Markov Model*, Krenn & Samuelsson [1996]) paramétereinek a becslése. A becsléshez rendelkezésre állnak a $P(\sigma_{k_1}\sigma_{k_2}...\sigma_{k_n})$-k empírikus közelítései, valamint felhasználhatjuk a jelek empírikus gyakoriságait. Az i-ik jel ν_i frekvenciájára adható elméleti becslés:

$$\nu_i = \sum_{j=1}^{N} v_j a_{ji},$$

hiszen v_j valószínűséggel van a rendszer az s_j állapotban, és ekkor a_{ji} valószínűséggel bocsátja ki a σ_i jelet.

Izgalmas kérdés, vajon mennyire szöveg-, téma- és nyelvspecifikusak a rejtett paraméterek, milyen mértékű eltérések engedélyezettek, milyen mértékű eltérések szükségesek a Damashek-módszer sikeréhez.

Befejezésül, még arról a kérdésről, hogy miért nem egyszerűsítettem le a gondolatmenetemet Markov-láncokra. Valószínűleg egyszerűbbek lennének a számítások, ha Markov-láncot, és nem Markov-modellt tekintünk. Két okból ragaszkodtam a Markov-modellhez. Egyrészt, a reguláris grammatikákat Markov-modellekre, és nem Markov-láncokra vezettük vissza. A második ok nyelvészeti: a fonotaktikai szabályok sok esetben felhasználnak metrikus fonológiai információkat is (szótagszerkezet, szószerkezet, stb.). Ezt úgy valósíthatjuk meg, ha a metrikus szerkezet elemeit (pl. szótagkezdet, mag, szótagzárlat) az állapotokkal reprezentáljuk, míg a szegmentumoknak (fonémáknak) felelnek meg a Markov-modell jelei. De ennek a kérdésnek a bővebb kidolgozása már egy nyelvészeti, és nem egy fizikai dolgozat tárgya lehetne.

7. Kódoló és nem-kódoló DNS-szakaszok

Marc Damashek cikke adta az ötletet, hogy a cikkben kidolgozott módszert DNS-szekvenciákra is alkalmazzam. [1] A DNS-szekvenciák statisztikai módszerekkel történő kutatásának jelenleg legizgalmasabb kérdése a kódoló és a nem-kódoló szakaszok eltérő viselkedése, ennek a magyarázatának a megtalálása, valamint olyan algoritmus készítése, amely automatikusan és nagy pontossággal választja szét a két típusú szekvenciákat. [2] Mantegna et al. [1994, 1995a], Czirók et al. [1995, 1996] bázispárokból képezett n-esekre végeztek Zipf-analízist (*n-tupple Zipf analysis*)($n = 3...8$), és azt kapták, hogy az előfordulási frekvencia a gyakorisági sorrendben elfoglalt hely függvényében a kódoló szakaszok esetén exponenciálisként, míg a nem-kódoló szakaszok esetén hatványfüggvényként cseng le (v. ö. az 5.1 fejezetben említett Zipf-törvénnyel).

Jogosan merül fel a kérdés, vajon megkülönböztethetők-e a kódoló és a nem-kódoló szakaszok a Damashek-féle szorzat segítségével. A 7.1 táblázat a HUMHBB szekvencia (humán hemoglobin) öt, átírásra kerülő szakaszából ("primary transcript") készített vektorokat tartalmaz. Egy ilyen szakasz áll egy "bevezető" részből, három exonból (kódoló szekvencia), a köztük lévő két intronból, valamint egy "záró" részből. Az E-vel jelölt vektorokat az exonok konkatenációiból képeztem, míg az I-vel jelölteket a fennmaradó (nem kódoló) részekből. (A táblázatban feltüntetett adatok esetében, a vektorok képzésénél $n = 3$ hosszúságú ablakot használtam, de más n-re is hasonló jelenségek figyelhetőek meg.)

[1] DNS-szekvenciák Markov-láncokkal történő elemzésére az 1980-as években is tettek már kísérletet, ld. Weir [1990] (pp. 237. skk.), ahol további irodalom is található.

[2] Ld. pl. Herzel & Große [1997], Große et al. [1997]

A táblázatban látványosan elkülönülnek a kódoló és a nem-kódoló szekvenciák. Az exonok egymás közötti szorzatai átlagosan $0,94 \pm 0,016$ értéket eredményeznek. A magas átlag és a kis szórás annál inkább meglepő, mert nagyon rövid szekvenciákról van szó. (Gondoljunk a *6.1* és *6.2* táblázatokra, ahol a többi szövegnél jóval rövidebb francia szövegek szorzása rosszabb eredményt adott, mint a hosszabb szövegeké.) A nem-kódoló szakaszok szintén magas átlagot eredményeznek: $0,92 \pm 0,03$. Ezzel szemben, kódoló és nem-kódoló szakaszból készített vektor szorzata szignifikánsan alacsonyabb, és nagy szórást mutat: $0,75 \pm 0,08$. Hasonló eredményeket figyelhetünk meg más n-ekre, valamint a RATCRYG-szekvenciára (norvég patkány) is.

Ezt követően, az egyes szekvenciákból képeztem egy "naív-vektort" is, amelynek i-ik komponensét az i-ik bázispár-n-es egyes bázispárjai előfordulási frekvenciája szorzataként számítottam ki ("korrelációmentes vektor"). Azaz, ha az i-ik n-gramm (bázis-n-es) alakja $\sigma_{k_1}\sigma_{k_2}...\sigma_{k_n}$, és ν_j jelöli a σ_j bázispár megfigyelt előfordulási frekvenciáját, akkor az $\mathbf{x}^{(naiv)}$ vektor i-ik komponense:

$$x_i^{(naiv)} := P^{(naiv)}(\sigma_{k_1}\sigma_{k_2}...\sigma_{k_n}) := \prod_{j=1}^{n} \nu_{k_j}.$$

A *7.2* táblázatban jól látszódik, hogy a vizsgált emlős szekvenciák esetében (a szekvenciákat ugyan úgy készítettem el, mint a *7.1* táblázat esetében) a kódoló szakaszok jóval távolabb esnek a korrelálatlan vektortól, mint a nem-kódolók: a szorzat értéke az előbbi esetben $0,66 \pm 0,05$, míg az utóbbiban $0,83 \pm 0,03$. Ez az eredmény nem mond ellent annak, hogy hosszútávú korrelációk az intront tartalmazó szekvenciákban fordulnak elő, míg a tisztán kódoló szakaszokban nem. Ugyanis, mint azt a 4. fejezetben kifejtettem, elképzelésem szerint az intront is tartalmazó szekvenciákat nem-reguláris folyamatok hozzák létre, és így születnek a hosszútávú korrelációk, míg az exonok most felfedezett rövidtávú korrelációit Markov-folyamattal is magyarázhatjuk.

A *7.2* táblázat eredményein felbuzdulva, a 6. fejezetben ismertett "mozgó doboz"-os módszert alkalmaztam a DNAS-szekvenciákra is. A HUMHBB átírásra kerülő részei (az exonok környékei) nagyon jellegzetes struktúrát mutatnak (*7.1 ábra*). A három exon közül, az első kettő nagyon közel van egymáshoz, így van olyan helyzet, hogy a doboz belelóg mindkettőbe. Ennek ellenére felfedezhető az a jellegzetes völgy, amely úgy keletkezik, hogy minél jobban nyúlik bele a doboz az exonba, annál alacsonyabb lesz a naiv-vektorral történő szorzás eredménye. A *primary transcript* első két exonja körüli völgy tehát átlapol, viszont a harmadik exon völgye szépen kivehető. Meglepő felfedezés ugyanakkor a harmadik exont megelőző völgy feltűnése, mind az öt átírásra kerülő szakasz esetén ugyan ott. Ezen "álexon" létére még nem találtam magyarázatot.

Végül, a *7.2 ábrán* az élesztő III. kromoszómájának (SCCHRIII) egy részlete látható. Ugyan ezt a részletet mutatja be Stanley et al. [1993b] 3. ábrája (a *DNA-walk* α-exponense egy $m = 800bp$ széles mozgó dobozban) Mindkét ábrában lokális minimumok jelzik az exonok helyét, ennek ellenére a két ábra kevés hasonlóságot mutat.

8. Befejezés

DNS-szekvenciák és természetes nyelvi szövegek esetén egyaránt felfedeztek az 1990-es évek elején hosszútávú korrelációkat. Ezt követően, az 1990-es évek közepén kiderült, hogy a DNS nemkódoló szakaszai és az írott szövegek további közös tulajdonságokkal rendelkeznek: például a karakter n-esekre készített Zipf-féle frekvencia−gyakoriság függvény a nem-kódoló DNS-szekvenciákra, akár csak a szintaxisra (szavak eloszlására egy szövegben) hatványfüggvényt követ, szemben a kódoló DNS-szakaszokkal és a betűk eloszlásával valamely szövegben (a fonológia leképeződése a helyesírásra), amelyek esetén a Zipf-függvény exponenciálisan cseng le (Mantegna et al. [1994, 1995a]). Ugyan ezek a vizsgálatok empírikusan hozták ki azt az eredményt, amit én a korrelációk meglétéből következtettem ki: a nemkódoló DNS-szekvenciákat nem lehet helyesen modellezni Markov-folyamatokkal (ellentétben az 1980-as évek próbálkozásaival, Weir [1990]). Magasabb rendű Markov-láncok javíthatnak a leírás pontosságán, de végleges megoldást szintén nem adhatnak (Mantegna et al. [1995a]). Talán azért nő a pontosság a rend növelésével, mert az adekváltabb leírást adó sztochasztikus környezetfüggetlen grammatikákat közelítik meg?

Dolgozatomban ezen jelenségek mögé kívántam tekinteni. A két szimbólumsorozat egyaránt összetett struktúrákat kódol lineáris formában, ezért a hosszútávú korrelációk − véleményem szerint − a kódolandó nemlinearitásának a megnyilvánulásai. A másodlagos folyamatok (fonológia, fehérje szintézis) viszont reguláris szabályokat alkalmaznak. Nem világos egyelőre, hogy miért tűnnek el a korrelációk a fehérje-szintézis során a kódoló sza-

kaszokban, holott a beszédprodukcióban megtalálhatóak egyaránt a szintaxis hosszútávú, és a fonológia rövidtávú korrelációi. Megjegyzem, hogy a Zipf-függvény viselkedése ugyanezt a kettősséget követi: a reguláris (korrelációmentes) szinten exponenciálisként cseng le, míg a korrelációkat produkáló szinten hatványfüggvénnyel közelíthető.

Dolgozatom végén a Damashek-féle vektortér/technika alkalmazását mutattam be – mint a reguláris folyamatok által létrehozott rövidtávú korrealációkon alapuló módszert – szövegeken, és – újdonságként – és DNS-en. DNS-ek esetében, a kódoló és a nem-kódoló szakaszok szorzatai szignifikánsan eltérnek egymástól, így egy újabb jelenséggel bővült a kódoló és nem-kódoló szakaszok statisztikai tulajdonságait magyarázni kívánó kutatók feladata. A felfedezés egyelőre nem alkalmas még a kétféle szekvencia automatikus

elválasztására, mivel találtam olyan szakaszokat is, amelyek exonként viselkednek, habár nem azok.

A természetes nyelvi szövegek kapcsán a módszer sikerességét helyesírási és fonotaktikai okokkal (nyelvek szerinti szortírozás), illetve a szöveg tartalmára jellemző szavak gyakoriságával (téma szerinti szétválogatás) magyaráztam. Ennek alátámasztására, javaslatot tettem a fonotaktika Markov-modellekkel történő megközelítésére.

Köszönetnyilvánítás

Köszönetemet fejezem ki *Vicsek Tamás* egyetemi tanárnak, aki felhívta a figyelmemet a statisztikus fizika nyelvészeti és genetikai vonatkozású fejezeteire, és minden téren támogatta a munkámat. Hálás vagyok *Czirók András*nak, az ELTE TTK Atomfizikai Tanszék doktoranduszának, aki cikkekkel és praktikus tanácsokkal segített, valamint *Rebrus Péter*nek, az MTA Nyelvtudományi Intézet PhD-ösztöndíjasának, aki a nyelvészeti vonatkozásokban volt a segítségemre.

Irodalomjegyzék

L. A. N. Amaral, S. V. Buldyrev, S. Havlin, M. A. Salinger, H. E. Stanley [1997]: Power law scaling in a system of interacting units with complex internal structure (preprint).

M. Amit, Y. Shmerler, E. Eisenberg, M. Abraham, N. Shnerb [1994]: Language and Codification Dependence of Long-Range Correlations in Texts, Fractals, **2**, 1, pp. 7-13.

S. Bird, T. M. Ellison [1994]: One-level Phonology: Autosegmental Representations and Rules as Finite Automata, Computational Linguistics, **20**, 1, pp. 55-90.

J-Ph. Bouchaud, M. Potters [1997]: Théorie des Risques Financiers. Portefeuilles, options et risques majeurs, Collection Aléa Saclay.

N. Chomsky [1957]: Syntactic Structres, The Hague: Mouton. Magyarul megjelent: Mondattani szerkezetek, Osiris-Századvég, Budapest, 1995.

A. Czirók, R. N. Mantegna, S. Havlin, H. E. Stanley [1995]: Correlations in binary sequences and a generalized Zipf analysis, Physical Review E, **52**, 1 , pp. 446-452.

A. Czirók, H. E. Stanley, T. Vicsek [1996]: Possible origin of power-law behavior in n-tuple Zipf analysis, Physical Review E **53**, 6371.

M. Damashek [1995]: Gauging Similarity with n-Grams: Language-Independent Categorization of Text, Science, **267**, pp. 843-848.

I. Derényi, T. Vicsek [1996]: The kinesin walk: A dynamic model with elastically coupled heads, Proc. Natl. Acad. Sci. USA, **93**, pp. 6775-6779.

G. Dietler, Y.-C. Zhang [1994]: Crossover from White Noise to Long Range Correlated Noise in DNA Sequences and Writings, Fractals, **2**, 4, pp. 473-479.

W. Ebeling, A. Neiman [1995]: Long-range correlations between letters and sentences in texts, Physica A 215, pp. 233-241.

F. Family, T. Vicsek [1991] (szerk.): Dynamics of Fractal Surfaces, World Scientific.

S. A. H. Geritz, J. A. J. Metz, É. Kisdi, G. Meszéna [1997]: Dynamics of Adaptation and Evolutionary Branching, Physical Review Letters, **78**, 10, pp. 2024-2027.

S. Ghashghaie, W. Breymann, J. Peinke, P. Talkner, Y. Dodge [1996]: Turbulent cascades in foreign exchange markets, Nature, **381**, 27 June 1996, pp. 767-770.

I. Große, H. Herzel, S. V. Buldyrev, H. E. Stanley [1997]: Mutual information of coding and noncoding DNA (preprint).

H. Herzel, I. Große [1997]: Correlations in DNA sequences: The role of protein coding segments, Physical Review E, **55**, 1, pp. 800-810.

I. Kanter, D. A. Kessler [1994]: Markov Process: Linguistics, Zipf's Law and Long-Range Correlations. (Submitted to PRL, Oct. 20, 1994).

R. M. Kaplan, M. Kay [1994]: Regular Models of Phonological Rule Systems, Computational Linguistics, **20**, 3, pp. 331-378.

Kenesei I. [1995] (szerk.): A Nyelv és nyelvek, Akadémiai Kiadó, Budapest.

Kiefer F. [1994] (szerk.): Strukturális magyar nyelvtan, 2. kötet: Fonológia, Akadémiai Kiadó, Budapest.

B. Krenn, Ch. Samuelsson [1996]: The Linguist's Guide to Statistics, forrás: http://coli.uni-sb.de/ christer.

Y. Liu, P. Cizeau, M. Meyer, C.-K. Peng, H. E. Stanley [1997]: Correlations in Economic Time Series (preprint submitted to Elsevier Science).

K. Martinás, M. Moreau [1995] (szerk.): Complex Systems in Natural and Economic Sciences, Proceedings of the Workshop Methods of Non-Equilibrium Processes...

R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H. E. Stanley [1994]: Linguistic Features of Noncoding Sequences, Physical Review Letters, **73**, 23, pp. 3169-3172.

R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H. E. Stanley [1995a]: Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics, Phys. Rev. E, **52**, 3, pp. 2939-2950.

R. N. Mantegna, H. E. Stanley [1995b]: Scaling behaviour in the dynamics of an economic index, Nature, **376**, 6 July 1995, pp. 46-49.

R. N. Mantegna, H. E. Stanley [1996]: Turbulence and financial markets, Nature, **383**, 17. October 1996, pp. 587-588.

R. N. Mantegna, H. E. Stanley: Stock market dynamics and turbulence, Parallel analysis of fluctuation phenomena, Physica A, 3357 (1997.).

Nagy Ferenc [1986]: Kvantitatív Nyelvészet, Tankönyvkiadó, Budapest.

C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H. E. Stanley [1992]: Long-range correlations in nucleotide sequences, Nature, **356**, pp. 168-170.

C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, H. E. Stanley [1993]: Finite-size effects on long-range correlations: Implications for analyzing DNA sequences, Physical Review E, **47**, 5, pp. 3730-3733.

M. Potters, R. Cont, J-Ph. Bouchaud [1997]: Financial markets as adaptive systems (preprint).

Révész Gy. [1979]: Bevezetés a formális nyelvek elméletébe, Akadémiai Kiadó, Budapest.

A. Schenkel, J. Zhang, Y.-C. Zhang [1993]: Long Range Correlation in Human Writings, Fractals, **1**, 1, pp. 47-57.

M. H. R. Stanley, L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, Ph. Maass, M. A. Salinger, H. E. Stanley [1996a]: Scaling behaviour in the growth of companies, Nature, **379**, 29 Febr. 1996, pp. 804-806.

M. H. R. Stanley, L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, Ph. Maass, M. A. Salinger, H. E. Stanley [1996b]: Can Statistical Physics Contribute to the Science of Economics?, Fractals, **4**, 3 (1996), pp. 415-425.

H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, J. M. Hausdorff, S. Havlin, J. Mietus, C.-K. Peng, F. Sciortino, M. Simons [1992]:Fractal landscapes in biological systems: Long-range correlations in DNA and interbeat heart intervals, Physica A, 191, 1-12.

H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons [1993a]: Long-range power-law correlations in condensed matter physics and biophysics, Physica A 200, 4-24.

H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, S. M. Ossadnik, C.-K. Peng, M. Simons [1993b]: Fractal Landscapes in Biological Systems, Fractals, **1**, 3, pp. 283-301.

T. Vicsek [1992]: Fractal Growth Phenomena (second edition), World Scientific.

B. S. Weir [1990]: Genetic Data Analysis, Methods for Discrete Population Genetic Data, Sinauer Associates, Inc. Publishers, Sunderland, Mass.

G. K. Zipf [1935]: The Psychobiology of Language, Houghton Mifflin, Boston.

G. K. Zipf [1949]: Human Behavior and the Principle of Least Effort, Addison-Wesley Press.